PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of

Yuji KOBAYASHI

Serial No.: 10/762,126 Group Art Unit:

Filed: January 21, 2004 Examiner:

For: INFORMATION SEARCHING APPARATUS AND METHOD, INFORMATION SEARCHING PROGRAM, AND STORAGE MEDIUM STORING THE INFORMATION SEARCHING PROGRAM

Certificate of Mailing

I hereby certify that this paper is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on:

Date: 02/05/04

By: _____

Marc A. Rossi

CLAIM FOR PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

The benefit of the filing date of the following prior foreign application filed in the following country is hereby requested for the above-identified application and the priority provided in 35 U.S.C. § 119 is hereby claimed:

JAPAN 2003 - 013428 January 22, 2003

In support of this claim, a certified copy of said original foreign application is filed herewith. It is requested that the file of this application be marked to indicate that the requirements of 35 U.S.C. 119 have been fulfilled and that the Patent and Trademark Office kindly acknowledge receipt of this document.

Respectfully submitted,

02/05/04
Date

Marc A. Rossi
Registration No. 31,923

Attorney Docket: CANO:116

AP-1654-US-PCT

CFW00180

US

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出願年月日
Date of Application: ２００３年　１月２２日

出願番号
Application Number: 特願２００３－０１３４２８

[ST. 10/C]: [ＪＰ２００３－０１３４２８]

出願人
Applicant(s): キヤノン株式会社

２００４年　１月１４日

特許庁長官
Commissioner,
Japan Patent Office 今井康夫

出証番号　出証特２００３－３１１１１０８

【書類名】　特許願
【整理番号】　224320
【提出日】　平成15年 1月22日
【あて先】　特許庁長官殿
【国際特許分類】　G06F　1/00
【発明の名称】　情報検索装置
【請求項の数】　2
【発明者】
　【住所又は居所】　東京都大田区下丸子３丁目３０番２号　キヤノン株式会社内
　【氏名】　小林　雄二
【特許出願人】
　【識別番号】　000001007
　【氏名又は名称】　キヤノン株式会社
　【代表者】　御手洗　冨士夫
【代理人】
　【識別番号】　100081880
　【弁理士】
　【氏名又は名称】　渡部　敏彦
　【電話番号】　03(3580)8464
【手数料の表示】
　【予納台帳番号】　007065
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9703713
【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　情報検索装置

【特許請求の範囲】

【請求項１】　文書から抽出された単語を前記文書と対応付けて文書検索用の索引情報として登録する索引情報登録手段と、該索引情報登録手段により登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索手段とを有する情報検索装置であって、

前記文書から未知単語を抽出する未知単語抽出手段と、

前記文書の種別を判別する文書種別判別手段と、

前記文書種別判別手段により判別された文書の種別に応じて、前記未知単語抽出手段により抽出された未知単語の前記索引情報登録手段による索引情報としての登録の可否を決定する登録可否決定手段とを有することを特徴とする情報検索装置。

【請求項２】　文書から単語を抽出する単語抽出手段と、

前記単語抽出手段により抽出された単語を前記文書と対応付けて、文書検索用の索引情報として登録する索引情報登録手段と、

前記索引情報登録手段により登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索手段と、

文字認識処理された文書に含まれている文字コードの誤りを校正する文字校正手段とを有し、

前記単語抽出手段は、前記文字校正手段により校正された後の文書から単語を抽出することを特徴とする情報検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術の分野】

本発明は、文字、画像等のマルチメディア情報を含んだ文書から所望の文書を検索する情報検索装置に関する。

【０００２】

【従来の技術】

従来、マルチメディア情報を含んだ文書から所望の文書を検索するため全文検索と呼ばれる手法を適用した情報検索装置が知られている。この装置では、例えば、所望の検索用情報を表す検索語あるいは文等を入力し、蓄積された文書群から、入力された語あるいは文と一致する語あるいは文を有している文書を得るようにしている。

【０００３】

また、文書画像であっても文書の内容に基づく情報検索を可能にするために、文書画像に含まれている文字画像部分を文字認識し、文字認識結果得られる文字情報をもとに全文検索を適用可能なように構成した装置が考案されている。

【０００４】

しかし、文字認識結果得られた文字コードを含んだ文書（以下、「文字認識処理済み文書」と称する）には、誤認識された文字が含まれる可能性があるため、文字認識を経ないテキスト等の文書と同様の手法によって全文検索を行なった場合、検索漏れの増大や、本来の文書内容と異なる文字との思いがけない一致による誤検索が生じることがある。

【０００５】

そのため、文字認識処理済み文書に全文検索を適用する前に、検索対象となる文字認識処理済み文書の誤認識箇所をひとつひとつユーザが目視して修正することが通常行われる。

【０００６】

そしてこの目視修正を不要とするために、下記特許文献１では、文字認識の候補となる複数の文字をその確からしさの推定値とともに用いることで、誤認識が含まれる文字認識処理済み文書であっても、複数の候補文字を選択して検索漏れを低減させる手法が開示されている。すなわち、複数の文字認識候補文字を含めて検索することで検索漏れを低減させることができる。

【０００７】

【特許文献１】

特許第２５８６３７２号

【０００８】

【発明が解決しようとする課題】

しかしながら、例えば、図５に示される誤認識文字列「モルール」に対して、たまたま誤認識文字列と一致してしまう「ルール」と照合することを回避できないため、文字列「ルール」で検索すると、本来は「モノレール」であるこの誤認識文字列「モルール」と一致し、誤検索されてしまう等のように、精度の低下を招く恐れがある。

【０００９】

また、文字認識により得られた文字コードのみで構成される文字認識処理済み文書の場合は、上記特許文献１の手法を用いたとしても、文字認識過程における他の候補文字の情報が必要であるため、その効果を期待できず、検索漏れの増大や誤検索の問題が残る。

【００１０】

一方で、単に文字単位の索引情報のみならず、実際に存在する単語との照合を行う形態素解析を行って、抽出された単語を文書検索用の索引情報として登録する、単語索引による情報検索装置も提案されている。この単語検索による情報検索装置では、文字単位の索引による情報検索装置と比べ、単語の境界をまたいだ一致等を避けることができるため、検索精度を向上させることができる。ところが、この単語索引による情報検索においても、すべての単語を単語辞書に収録することは事実上不可能であるため、辞書に存在しない単語について検索することができず、検索漏れが起こり得る。

【００１１】

本発明は上記従来技術の問題を解決するためになされたものであり、その第１の目的は、検索漏れ及び誤検索の少ない適切な検索を行うことができる情報検索装置を提供することにある。

【００１２】

本発明の第２の目的は、検索精度を向上させることができる情報検索装置を提供することにある。

【００１３】

【課題を解決するための手段】

上記第１の目的を達成するために本発明の請求項１の情報検索装置は、文書から抽出された単語を前記文書と対応付けて文書検索用の索引情報として登録する索引情報登録手段と、該索引情報登録手段により登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索手段とを有する情報検索装置であって、前記文書から未知単語を抽出する未知単語抽出手段と、前記文書の種別を判別する文書種別判別手段と、前記文書種別判別手段により判別された文書の種別に応じて、前記未知単語抽出手段により抽出された未知単語の前記索引情報登録手段による索引情報としての登録の可否を決定する登録可否決定手段とを有することを特徴とする。

【００１４】

上記第２の目的を達成するために本発明の請求項２の情報検索装置は、文書から単語を抽出する単語抽出手段と、前記単語抽出手段により抽出された単語を前記文書と対応付けて、文書検索用の索引情報として登録する索引情報登録手段と、前記索引情報登録手段により登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索手段と、文字認識処理された文書に含まれている文字コードの誤りを校正する文字校正手段とを有し、前記単語抽出手段は、前記文字校正手段により校正された後の文書から単語を抽出することを特徴とする。

【００１５】

【発明の実施の形態】

以下、本発明の実施の形態を図面を参照して説明する。

【００１６】

図１は、本発明の一実施の形態に係る情報検索装置の全体構成を示すブロック図である。

【００１７】

図１において、１１はマイクロプロセッサ（ＣＰＵ）（索引情報登録手段、文書検索手段、未知単語抽出手段、単語抽出手段、文書種別判別手段、文字認識処理手段、登録可否決定手段、文字校正手段）であり、情報検索のための演算、論理判断等を行い、アドレスバスＡＢ、コントロールバスＣＢ、データバスＤＢを

介して、それらのバスに接続された各構成要素を制御する。

【００１８】

アドレスバスＡＢはＣＰＵ１１が制御の対象とする構成要素を指示するアドレス信号を転送する。コントロールバスＣＢはＣＰＵ１１が制御の対象とする各構成要素のコントロール信号を転送して印加する。データバスＤＢは各構成要素相互間のデータ転送を行う。

【００１９】

１２は読出し専用の固定メモリ（ＲＯＭ）であり、本実施の形態で実行される処理プログラム等の制御プログラムコードを記憶する。１３は書込み可能なランダムアクセスメモリ（ＲＡＭ）であり、各構成要素からの各種データの一時記憶に用いられる。ＲＡＭ１３はまた、図２において後述する検索語保持部２０２、検索結果保持部２０４、未知語検索指定保持部２０８（可否指定手段）、文書タイプ保持部２１４を記憶する。

【００２０】

１４は外部メモリ（ＤＩＳＫ）であり、図２において後述する単語インデックス２０５、単語抽出辞書２０７（単語辞書）、文字校正用辞書２１０、抽出テキスト２１２、文書登録処理部２１５に入力された登録文書が格納される。また、図２において後述する検索要求入力処理部２０１、検索処理部２０３（文書検索手段）、単語抽出処理部２０６（未知単語抽出手段、単語抽出手段）、文字校正処理部２０９（文字校正手段）、文字認識処理部２１１（文字認識処理手段）、登録文書タイプ判定部２１３（文書種別判別手段）の各処理部を実行するプログラムコードが格納される。また、これらのデータ及びプログラムを格納する記憶媒体としては、固定磁気ディスク、ＲＯＭ、フロッピー（登録商標）ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ、ＤＶＤ－ＲＯＭ、ＤＶＤ－ＲＡＭ、ＤＶＤ－Ｒ、メモリカード、光磁気ディスク、磁気テープ等を用いることができる。

【００２１】

１５はキーボード（ＫＢ）であり、アルファベットキー、ひらがなキー、カタカナキー、句点等の文字記号入力キー、検索を指示する検索キー及び、カーソル移動を指示するカーソル移動キー等のような各種の機能キーを備えている。１６

は表示用ビデオメモリ（ＶＲＡＭ）であり、表示すべきデータのパターンを蓄える。１７はＣＲＴコントローラ（ＣＲＴＣ）であり、表示用ビデオメモリＶＲＡＭ１６に蓄えられた内容を後述する表示器ＣＲＴ１８に表示する役割を担う。

【００２２】

表示器ＣＲＴ１８は、陰極線管、液晶パネル等の表示装置であり、表示装置ＣＲＴ１８におけるドット構成の表示パターン及びカーソルの表示は、ＣＲＴコントローラ１７により制御される。１９はネットワークコントローラ（ＮＩＣ）であり、Ethernet（登録商標）等のネットワークに接続する役割を担う。

【００２３】

かかる各構成要素からなる情報検索装置は、キーボードＫＢ１５からの各種の入力及びネットワークコントローラ１９から供給されるネットワーク経由の各種入力に応じて作動し、キーボードＫＢ１５からの入力及びネットワークコントローラ１９からの入力が供給されると、まず、インタラプト信号がＣＰＵ１１に送られ、ＣＰＵ１１がＤＩＳＫ１４内に記憶されている各種の制御信号を読み出し、それらの制御信号に従って、各種の制御を行う。

【００２４】

図２は、本発明の実施の形態の情報検索装置の機能構成を示すブロック図である。

【００２５】

同図において、２０１は所望の検索対象に関する要求事項（検索文や検索語等の検索用情報であって、本実施の形態では、検索語を用いることとする）を入力する検索要求入力処理部である。２０２は検索要求入力処理部２０１によって入力された検索語を記憶する検索語保持部である。２０３は検索語保持部２０２に記憶された検索語に基づいて、登録された文書を対象として検索する検索処理部である。２０４は、検索処理部２０３の処理結果を保持する検索結果保持部である。

【００２６】

２０５は、文書登録処理部２１５で登録された文書から抽出された単語と、その単語の出現する文書の文書情報とを対応させて記憶した単語インデックスであ

る（後述）。２０６は、文書登録処理部２１５で登録された文書から検索対象とする単語（検索語との照合対象となる単語）を抽出する単語抽出処理部である。２０７は、単語抽出処理部２０６で参照される単語を定義した単語抽出辞書である。

【００２７】

２０８は、文書登録処理部２１５で登録された文書において出現した単語のうち単語抽出辞書２０７において定義されていない「未知語」であると判定された単語に対して、これを単語インデックス２０５に、文書検索用の索引情報として登録する否かを指定する未知語検索指定保持部である。この未知語検索指定保持部２０８では、文字認識処理を経た結果であるテキスト文書（すなわち、文字認識の結果得られた文字コードを含んだ文書であり、以下、「文字認識処理済み文書」とも称する）の登録の際に、単語抽出辞書２０７に収録されていない未知語であっても検索可能なように登録する（インデックスを作成する）か、あるいは未知語については検索対象とならないように登録を禁止するかが、ユーザの意思で指定される。すなわち、未知語検索指定保持部２０８は、文字認識処理済み文書から抽出された未知語を索引情報として登録することを許可するか否かの指定を記憶するものである。

【００２８】

２０９は、文字認識処理部２１１において文書画像から得られた認識文字よりなる抽出テキスト２１２に対して、文字校正用辞書２１０及び単語抽出辞書２０７を参照して文字認識の誤りを校正する文字校正処理部である。２１１は、文書登録処理部２１５で登録された文書が画像文書である場合に、その画像情報を文字認識処理により文字情報に変換する文字認識処理部である。２１２は、文字認識処理部２１１の処理結果である認識文字（文字認識処理済み文書を構成する文字コード）を格納する抽出テキストである。

【００２９】

２１３は、文書登録処理部２１５で登録された文書の文書タイプ（画像文書であるか、文字認識結果を格納した抽出テキストであるか、一般テキストであるか）を判定する登録文書タイプ判定部である。登録文書タイプ判定部２１３では、

与えられた登録文書のファイル名の拡張子が"bmp"、"jpg"、"gif"、"tif"等の画像フォーマットをあらわす場合には、文書タイプ保持部２１４に「画像タイプ」であることを格納し、登録文書のファイル名の拡張子が"txt"等、テキストをあらわす場合には、文書タイプ保持部２１４に「テキストタイプ」であることを格納する。また、テキスト文書でも特に文字認識処理の結果得られた抽出テキストである場合には、特別な拡張子として"ocr"を使用するものとし、拡張子が"ocr"であった場合は、文書タイプ保持部２１４に「認識結果テキストタイプ」であることを格納する。２１５は文書の登録を行う文書登録処理部である。

【００３０】

次に、検索要求入力処理部２０１において、文書検索の検索要求のための検索語を操作者が指示する場合における操作パネルの表示例について、図３を用いて説明する。

【００３１】

図３は、検索語を指示する場合において、表示器ＣＲＴ１８に画面表示される操作パネルの例を示す図である。

【００３２】

図３において、３０１は検索要求入力操作を行う表示ウインドウである。３０２は検索語等の検索用情報を入力する検索文入力領域である。３０３は入力中の検索用情報（検索要求文）の一例を示しており、同図においては、「カラープリンタの売上」と入力されている。３０４は検索文入力領域３０２における入力位置を示す入力カーソルである。

【００３３】

３０８は検索処理の実行を指定する検索実行ボタンであり、検索実行ボタン３０８を押下することで、指定した検索処理が実行される。３０９は検索処理の終了あるいは中止を指定するキャンセルボタンであり、キャンセルボタン３０９を押下すると、ただちに検索処理が終了し、表示ウインドウ３０１が閉じられて終了する。３１０は検索ボタン３０８の押下によって検索処理を行った結果を表示する検索結果表示領域であり、同図においては検索処理がなされていない状態であるので、何も表示されていない。

【００３４】

図４は、検索要求入力処理部２０１による検索処理実行後の検索結果の表示例を示す図である。同図には、検索用情報３０３に示された「カラープリンタの売上」に対して、文書検索を行った検索結果が表示されている。

【００３５】

図４において、４０２は、検索結果の順位を示すランク表示領域である。検索結果は検索要求に適合している順にランク付けされ、ランク順に表示される。図４の表示例においては、ランク２５位から３０位までの検索結果が表示されている。４０３は、検索された文書の表題であり、４０４は文書のファイル名である。４０５は、検索された文書の大意が掴める程度の内容を表示する文書内容表示である。文書内容表示４０５には、あらかじめ文書の書誌的属性として与えられた要約文、あるいは文書内容を自動的に要約した要約文、あるいは文書の一部を大意として抽出した大意文等を表示することができる。

【００３６】

４０６は、表示位置を指定するために同種のウインドウ表示装置において用いられているエレベータバーであり、検索結果表示領域３１０に表示しきれない場合に、検索結果表示領域３１０内において検索結果を部分表示しながら、表示されていない他の部分を表示するために用いられる。

【００３７】

次に、文字認識処理を施した認識結果である抽出テキスト２１２について、図５を用いて説明する。

【００３８】

図５は、文字を表す画像に対して文字認識処理を施した結果得られた抽出テキストを示す図である。抽出テキストは、文字認識処理部２１１において抽出される場合と、あらかじめ文字認識処理が施されて文書登録処理部２１５に与えられている場合がありえる。

【００３９】

一般に、文字認識処理においては、認識誤りを含むことがありえる。例えば、同図に示すように、「モノク口」の４文字目「口」がカタカナの「ロ」（ろ）で

あるべきところ、漢字の「口」（くち）と誤認識され、文字認識処理の対象とされた元画像では「モノレール」と書かれていた部分が「モルール」と誤認識されている。

【００４０】

次に、文字認識結果である抽出テキストに対して、文字校正処理を行った処理結果の例について図６を用いて説明する。

【００４１】

図６は、図５に例示された文字認識誤りを含んでいる抽出テキストに対して文字校正処理を行った場合の結果を示す図である。文字校正処理部２０９では、後述する文字校正用辞書２１０と単語抽出辞書２０７を参照することにより、図５における誤認識である「モノク口」（「口」は漢字の「くち」）を「モノクロ」（「ロ」はカタカナ）と校正している。

【００４２】

図７は、文字校正用辞書２１０の構成を示す概念図である。文字校正用辞書２１０は、文字認識処理において誤認識の発生しやすい、似通った字形の文字を対応付けて格納したものである。同図において、１つの行に表されている文字同士が、互いに誤認識されやすいことを表している。

【００４３】

図８は、単語抽出辞書２０７の構成を示す概念図である。同図において、８０１は単語の見出し語、８０２は単語の品詞を示す。品詞８０２は、抽出された単語と前置される単語との接続可能性を判定するために用いられる。

【００４４】

図９は、単語インデックス２０５の構成を示す概念図である。単語インデックス２０５は、登録文書中に出現するすべての単語について、文書中における出現頻度とその文書との対応をとって格納するテーブルである。単語インデックス２０５におけるテーブルの第１列情報である９０１は、検索見出しとなる単語（インデックス見出し語）である。テーブルの第２列情報は、インデックス見出し語９０１が出現する文書と、その文書においてインデックス見出し語９０１が何回出現するかを対応づけた文書情報であり、インデックス見出し語９０１が出現す

るすべての文書について格納される。その際、文書は一意化された文書識別番号で記録される。

【００４５】

例えば、インデックス見出し語９０１「カラー」に対応して、文書情報９０２において（１０００、１５）、（１２００、５）等と記録されており、これによれば、「カラー」は、文書識別番号「１０００」で表される文書に１５回出現し、文書識別番号「１２００」で表される文書に５回出現していることが示される。

【００４６】

図１０は、文字校正処理で用いられる文字候補ラティスの構成を示す概念図である。

【００４７】

文字候補ラティスは、文字校正処理部２０９での文字校正処理において、誤認識の発生する可能性のある文字について、他の候補文字を格子状に配したものである。文字校正処理対象の抽出テキストの各文字について、図７に示す文字校正用辞書２１０に格納された文字のいずれかと一致したならば、対応している他の文字が置換候補文字として追加される。

【００４８】

図１０においては、図５に示す文字認識結果の抽出テキストに対して作成された文字候補ラティスの一部が例示されており、「このモノク口の写真に写っているモルールは当時の重要な交通手段であった」という文字列に対して、図７の文字校正用辞書２１０に出現する「口」と「－」に対して、候補文字を追加したラティスが作成されている。

【００４９】

図１１は、本実施の形態における文書検索処理のフローチャートを示す図である。

【００５０】

まず、ステップＳ１００１では、検索要求入力処理部２０１の動作を行う処理モジュールによって、検索要求入力処理を行う。検索要求入力処理では、図３に

示す操作パネル中の検索文入力領域３０２に入力された検索用情報（本実施の形態では検索語）を取り出し、単語抽出辞書２０７を参照して、入力検索用情報から単語を抽出し、抽出した検索語を検索語保持部２０２に格納する。

【００５１】

次に、ステップＳ１００２では、検索語保持部２０２に格納された検索語が含まれる文書を単語インデックス２０５を参照して検索する。すなわち、検索語保持部２０２より検索語を取り出し、取り出された検索語と一致するインデックス見出し語９０１を検索し、見つかった見出し語に対応する文書情報９０２を取り出す。そして、取り出された文書情報９０２中の文書識別番号と出現頻度とを検索結果保持部２０４に格納する。なお、既に同一の文書識別番号をもつ文書情報が検索結果保持部２０４に格納済みであった場合は、その出現頻度を更新する。検索語保持部２０２に格納されたすべての検索語についてこの処理を行い、処理が終わったならば、検索語保持部２０２に記憶された検索結果を出現頻度の大きい順にソートする。

【００５２】

次に、ステップＳ１００３では、前記ステップＳ１００２で検索された検索結果を、検索結果保持部２０４より取り出して表示する。なお、この処理は同種の情報検索装置において広く行われている公知の処理と同様になされる。その後、本処理を終了する。

【００５３】

図１２は、本実施の形態における文書登録処理のフローチャートを示す図である。

【００５４】

まず、ステップＳ３００１では、文書登録処理部２１５に入力された文書の文書タイプを判定する。文書タイプの判定は、文書登録処理部２１５に入力された文書ファイル名の拡張子で行い、拡張子が"bmp"、"jpg"、"gif"、"tif"等である場合は、「文書画像」であるので、文書タイプ保持部２１４に「画像タイプ」であることを格納するとともに、ステップＳ３００２へ進む。また、拡張子が"ocr"であれば、文字認識処理の結果得られた抽出テキストであるので、文書タ

イプ保持部２１４に「文字認識抽出テキストタイプ」であることを格納すると共に、抽出テキスト２１２にその内容を格納して、ステップＳ３００３へ進む。一方、拡張子が"txt"や"html"等である場合は、一般テキストであるので、文書タイプ保持部２１４に「テキストタイプ」であることを格納するとともに、ステップＳ３００４へ進む。

【００５５】

ステップＳ３００２では、画像タイプと判定された文書に対して、その画像中の文字部分について文字認識処理を行う。この場合、文書タイプが画像タイプから文字認識抽出テキストタイプに変わったので、認識文字の抽出テキストを抽出テキスト２１２に作成し、文書タイプ保持部２１４に「文字認識抽出テキストタイプ」であることを格納して、ステップＳ３００３に進む。なお、画像情報から文字画像と照合を行い、文字コード化を行う文字認識処理は公知の手法でなされる。

【００５６】

ステップＳ３００３では、抽出テキスト２１２に対して、文字認識の誤り訂正を行うために、後述する図１３の文字校正処理を行う。ここでは、前記ステップＳ３００２の文字認識処理の結果得られた認識文字抽出テキスト、あるいは前記ステップＳ３００１において、「文字認識抽出テキストタイプ」と判定された文書である抽出テキストに対して文字校正処理がなされる。

【００５７】

続くステップＳ３００４では、後述する図１４の単語抽出処理を実行する。すなわち、文字認識結果ではない一般の文書（一般テキスト）、及び前記ステップＳ３００３において文字校正処理の施された文書（文字認識抽出テキスト）から、単語を抽出し、文書検索のための単語インデックス２０６を作成する。

【００５８】

図１３は、図１２のステップＳ３００３で実行される文字校正処理のフローチャートを示す図である。

【００５９】

まず、ステップＳ２００１では、文書校正処理で用いる文字候補ラティスを作

成する。文字候補ラティスは、前述のように、文字認識結果の抽出テキストのうち、誤認識の起こりやすい文字について候補文字を加えて、ラティス上に構成したものである（図１０参照）。文字候補ラティス作成処理では、対象となる抽出テキスト２１２より文字を取り出し、文字校正用辞書２１０を参照し、取り出した文字が文字校正用辞書２１０に登録されている文字であった場合は、その文字を含む、似た字形のグループの各文字を候補文字として文字候補ラティスに追加する。その際、ラティスの各格子点の候補文字の１番目の文字としては元の抽出テキストの文字を配するものとする。

【００６０】

次に、ステップＳ２００２では、文字候補ラティスの各格子点の１番目の文字に従って単語抽出辞書２０７を検索する。すなわち、ラティスを構成する各格子に位置する１番目の文字列と一致する見出し語があるかどうかを検索する。

【００６１】

次に、ステップＳ２００３では、前記ステップＳ２００２での単語検索の結果、単語が検索されたかどうか判別する。単語が検索された場合は、ステップＳ２００８へ進む一方、単語が検索されなかった場合は、ステップＳ２００４へ進む。

【００６２】

ステップＳ２００４では、文字候補ラティスにまだ単語検索していない候補文字列があるかを判別し、未検索の候補文字列がなかったならば前記ステップＳ２００８へ進む一方、未検索の候補文字列があったならばステップＳ２００５へ進む。

【００６３】

ステップＳ２００５では、文字候補を変更して、未検索の候補文字列の単語検索を行う。次に、ステップＳ２００６では、前記ステップＳ２００５での単語検索の結果、一致する単語が検索されたか否かを判別し、単語が検索されなかったならば、再び前記ステップＳ２００４へ戻って、未検索の候補文字列があるかの判別を繰り返す。

【００６４】

一方、単語が検索されたならばステップＳ２００７へ進んで、前記ステップＳ２００５において検索された単語と一致する候補文字列を構成する文字に、抽出テキストの文字を置き換える。例えば、図１０の文字候補ラティスに対して、図８の単語抽出辞書２０７を検索した場合、候補文字列「モノクロ」と一致するため、抽出テキスト中の「モノク口」の「口」（漢字の「口」）を「ロ」（カタカナの「ロ」）に置換する。

【００６５】

次に、ステップＳ２００８では、次の単語検索位置を取得する。すなわち、単語抽出辞書２０７との照合の終わった文字候補列をスキップし、未照合の候補文字列開始位置を取得する。ここで、前記ステップＳ２００４から分岐してきた場合は、文字候補ラティス中の候補文字列と単語との照合がとれなかった場合であるので、最後に照合のとれた文字位置以降の最初の助詞の次の位置を、次の候補文字列開始位置とする。

【００６６】

次に、ステップＳ２００９では、前記ステップＳ２００８で得られた候補文字列開始位置が文書末尾に達したか、すなわち、抽出テキストのすべての文字列について単語抽出辞書２０７との照合による文字校正処理を終えたかどうかを判別し、文書末尾に達していなければ前記ステップＳ２００２へ戻って未処理の文字列について処理を繰り返す一方、文書末尾に達していたならば本処理を終了する。

【００６７】

図１４は、図１２のステップＳ３００４で実行される単語抽出処理のフローチャートを示す図である。

【００６８】

まず、ステップＳ４００１では、単語検索、すなわち、単語抽出処理対象のテキストの文字列と、単語抽出辞書２０７との照合を行う。そして、ステップＳ４００２では、単語が検索されたか否かを判別する。その判別の結果、単語が検索されなかった場合は、ステップＳ４００４に進む一方、単語が検索された場合は、ステップＳ４００３へ進む。

【００６９】

ステップＳ４００３では、検索された単語と、既に抽出済みの直前の単語（前接語）が接続可能か否かを、単語抽出辞書２０７の品詞情報に基づき接続判定表（不図示）を用いて判別する。前接語との接続可能性の判定手法、及び接続可能性判定に用いる接続判定表の構成については公知であるため、詳細な説明を省略する。

【００７０】

その判別の結果、抽出済みの単語と検索された単語とが接続可能である場合は、ステップＳ４００７へ進む一方、接続不可能である場合は、ステップＳ４００４へ進む。

【００７１】

ステップＳ４００４では、前記ステップＳ４００１での単語検索の結果、単語が検索されなかったか、あるいは、前接語と接続可能な単語が検索されなかった場合であるので、その検索の照合開始位置から、辞書登録されていない未知語の抽出処理を行う。

【００７２】

未知語の抽出処理は、例えば、連続するカタカナをひとつの未知語として抽出する、あるいは、照合開始位置の文字から頻度の高い助詞の出現する直前の文字までをひとつの未知語として抽出する等の公知の手法を適用可能である。この未知語抽出手法を、図６に示される文字校正処理を行ったテキストに対して適用することで、図８に示される単語抽出辞書２０７には未登録である「モルール」が未知語として抽出される。なお、上述のように、図６における「モルール」は、本来は「モノレール」であるべきところ、誤認識されたものである。

【００７３】

次に、ステップＳ４００５では、単語抽出処理対象文書が文字認識結果文書、すなわち、文字認識処理済み文書であるか否かを判定する。文字認識結果文書であるかどうかは文書タイプ保持部２１４に記憶された文書タイプを参照することにより行われる。例えば、図１２のステップＳ３００１またはステップＳ３００２において、文書タイプ保持部２１４に「文字認識抽出テキストタイプ」と記憶

されていれば、文字認識結果文書であると判別される。

【００７４】

その判別の結果、単語抽出処理対象文書が文字認識結果文書（文字認識処理済み文書）でない場合は、ステップＳ４００７に進む一方、単語抽出処理対象文書が文字認識結果文書（文字認識処理済み文書）である場合は、ステップＳ４００６へ進み、文字認識結果文書に対する未知語検索指定の有無を判定する。未知語検索指定の有無の判別は、未知語検索指定保持部２０８を参照することにより行われる。ここで、「未知語検索指定有り」は、文字認識処理済み文書から抽出された未知語を単語インデックス２０５に索引情報として登録することを許可する指定であり、「未知語検索指定無し」は禁止する指定である。

【００７５】

その判別の結果、「未知語検索指定有り」である場合は、前記ステップＳ４００７へ進む。

【００７６】

ステップＳ４００７では、前記ステップＳ４００１で検索された（既知の）単語、あるいは前記ステップＳ４００４で抽出された未知語を、単語インデックス２０５に索引情報として登録する。その際、登録しようとする単語が単語インデックス２０５のインデックス見出し語９０１に既に存在している場合において、インデックス見出し語９０１に対応する文書情報９０２に、当該文書の文書識別番号が存在するときは、その文書識別番号に対応する出現回数を１だけ加算する。また、登録しようとする単語が単語インデックス２０５のインデックス見出し語９０１に既に存在している場合において、インデックス見出し語９０１に対応する文書情報９０２に、当該文書の文書識別番号が存在しないときは、抽出テキストの文書識別番号を新たに登録し、出現回数を１とする。その後、ステップＳ４００８に進む。

【００７７】

一方、前記ステップＳ４００６の判別の結果、「未知語検索指定無し」である場合は、抽出された未知語を単語インデックス２０５に索引情報として登録することなく、ステップＳ４００８へ進む。

【００７８】

ステップＳ４００８では、単語抽出処理対象の文書（テキスト）のすべての文字列に対する処理を終えたか否かを、文書末尾に達したかどうかで判別する。その判別の結果、文書末尾に達していなければ前記ステップＳ４００１へ戻って、未処理の文字列について上述の処理を繰り返す一方、文書末尾に達していれば本処理を終了する。

【００７９】

本実施の形態によれば、文字認識処理済み文書から抽出した未知語について、単語インデックス２０５に索引情報として登録するか否かを、未知語検索指定保持部２０８で指定できるようにしたので、検索漏れの抑制と検索精度の向上のいずれに重点を置くかという、検索者の意図を反映させて未知語の索引登録可否を決定することで、使い勝手がよく、検索漏れ及び誤検索の少ない適切な検索を行うことができる。

【００８０】

また、文書が文字認識処理済み文書でない場合、該文書から抽出された単語については、索引情報としての登録を一律に許可するようにしたので、高速な索引登録処理を実現することができる。

【００８１】

また、文字認識処理を経た文書について単語抽出処理を行う場合は、事前に文字校正処理を施すようにしたので、情報検索の精度を向上させることができる。

【００８２】

なお、本実施の形態では、文字認識処理済み文書から抽出した未知語については、「未知語検索指定有り」の場合にのみ、単語インデックス２０５に索引情報として登録されるようにしたが（図１４のステップＳ４００６、Ｓ４００７）、これに限るものでなく、文字認識処理済み文書から抽出した未知語については、一律に、単語インデックス２０５への索引情報としての登録を禁止するように構成してもよい。その場合は、図１４のステップＳ４００６の処理を省略すると共に、ステップＳ４００５の判別の結果、単語抽出処理対象文書が文字認識結果文書（文字認識処理済み文書）である場合は、直ちに前記ステップＳ４００８に進

むように処理すればよい。

【００８３】

このようにすれば、単語抽出の対象となった文書の種別によって、単語インデックス２０５への索引情報としての登録の可否が決定されるので、例えば、文字認識処理済み文書から抽出した未知語については無駄で不適切な索引登録を抑制でき、索引登録処理時間の短縮化、索引サイズの縮小という効果が得られる。また、文字認識誤りの含まれ得る文書であっても、認識誤りに由来する不適切な索引登録を回避して、不適切な検索結果を抑制することができ、操作性に優れた、精度の高い情報検索が可能となる。一方では、文字認識誤りのない文書から抽出された単語については一律に登録可能とすることで、高速な索引登録処理を実現することができる。よって、文書の種別に応じて未知単語の索引登録の可否を決定することで、検索漏れ及び誤検索の少ない適切な検索を行うことができる。

【００８４】

なお、本実施の形態では、単語インデックス２０５、単語抽出辞書２０７、文字校正用辞書２１０を単一の装置を構成するＤＩＳＫ１４に配置するものとして説明したが、これらの構成要素を異なる装置に分散配置し、ＮＩＣ１９を介してネットワーク上で処理を行うようにすることも可能である。

【００８５】

なお、本発明は、複数の機器（例えばホストコンピュータ、インタフェース機器、リーダ、プリンタ等）から構成されるシステムに適用しても、ひとつの機器からなる装置（例えば、複写機、ファクシミリ装置等）に適用してもよい。

【００８６】

また、本発明の目的は、実施の形態の機能を実現するソフトウェアのプログラムコードを記録した記憶媒体を、システム或いは装置に供給し、そのシステム或いは装置のコンピュータ（またはＣＰＵやＭＰＵ等）が記憶媒体に格納されたプログラムコードを読み出して実行することによっても達成される。

【００８７】

この場合、記憶媒体から読み出されたプログラムコード自体が前述した実施の形態の機能を実現することになり、そのプログラムコードを記憶した記憶媒体は

本発明を構成することになる。

【００８８】

又、プログラムコードを供給するための記憶媒体としては、例えば、フロッピー（登録商標）ディスク、ハードディスク、光磁気ディスク、ＣＤ－ＲＯＭ、ＣＤ－Ｒ、ＣＤ－ＲＷ、ＤＶＤ－ＲＯＭ、ＤＶＤ－ＲＡＭ、ＤＶＤ－ＲＷ、ＤＶＤ＋ＲＷ、磁気テープ、不揮発性のメモリカード、ＲＯＭ等を用いることができる。

【００８９】

また、コンピュータが読み出したプログラムコードを実行することにより、上記実施の形態の機能が実現されるだけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼動しているＯＳ（オペレーティングシステム）等が実際の処理の一部または全部を行い、その処理によって前述した実施の形態の機能が実現される場合も含まれる。

【００９０】

更に、記憶媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるＣＰＵ等が実際の処理の一部または全部を行い、その処理によって前述した実施の形態の機能が実現される場合も含まれる。

【００９１】

本発明の様々な例と実施例が示され説明されたが、当業者であれば、本発明の趣旨と範囲は本明細書の特定の説明と図に限定されるのではなく、本願特許請求の範囲にすべて述べられた様々の修正と変更に及ぶことが理解されるであろう。

【００９２】

本発明の実施態様の例を以下に列挙する。

【００９３】

〔実施態様１〕　文書から抽出された単語を前記文書と対応付けて文書検索用の索引情報として登録する索引情報登録手段と、該索引情報登録手段により登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文

書検索手段とを有する情報検索装置であって、前記文書から未知単語を抽出する未知単語抽出手段と、前記文書の種別を判別する文書種別判別手段と、前記文書種別判別手段により判別された文書の種別に応じて、前記未知単語抽出手段により抽出された未知単語の前記索引情報登録手段による索引情報としての登録の可否を決定する登録可否決定手段とを有することを特徴とする情報検索装置。

【００９４】

〔実施態様２〕　前記文書種別判別手段は、前記文書が文字認識処理により得られた文字コードを含んだ文字認識処理済み文書であるか否かを判別し、前記登録可否決定手段は、前記文書が前記文字認識処理済み文書である場合は、該文書から前記未知単語抽出手段により抽出された未知単語の前記索引情報登録手段による索引情報としての登録を禁止することを特徴とする実施態様１記載の情報検索装置。

【００９５】

〔実施態様３〕　前記文書種別判別手段は、前記文書が文字認識処理により得られた文字コードを含んだ文字認識処理済み文書であるか否かを判別するものであり、さらに、前記文書が前記文字認識処理済み文書である場合は、該文書から前記未知単語抽出手段により抽出された未知単語の前記索引情報登録手段による索引情報としての登録可否を指定する可否指定手段を有し、前記登録可否決定手段は、前記文書が前記文字認識処理済み文書である場合は、前記可否指定手段の指定に基づいて、前記未知単語抽出手段により抽出された未知単語の前記索引情報登録手段による索引情報としての登録の可否を決定することを特徴とする実施態様１記載の情報検索装置。

【００９６】

〔実施態様４〕　前記文書種別判別手段は、前記文書が文字認識処理により得られた文字コードを含んだ文字認識処理済み文書であるか否かを判別し、前記登録可否決定手段は、前記文書が前記文字認識処理済み文書でない場合は、該文書から前記未知単語抽出手段により抽出された未知単語の前記索引情報登録手段による索引情報としての登録を許可する実施態様１～３のいずれかに記載の情報検索装置。

【００９７】

〔実施態様５〕　前記文書種別判別手段は、前記文書が文字認識処理により得られた文字コードを含んだ文字認識処理済み文書であるか否かを判別するものであり、さらに、前記文書が前記文字認識処理済み文書である場合は、それに含まれている文字コードの誤りを校正する文字校正手段を有し、前記未知単語抽出手段は、前記文字校正手段により校正された後の文書から未知単語を抽出することを特徴とする実施態様１～４のいずれかに記載の情報検索装置。

【００９８】

〔実施態様６〕　前記文書種別判別手段は、前記文書が画像文書であるか否かを判別するものであり、さらに、前記文書が画像文書である場合は、該画像文書から文字認識処理により文字コードを得る文字認識処理手段と、該文字認識処理手段により文字認識処理された文書に含まれている文字コードの誤りを校正する文字校正手段とを有し、前記未知単語抽出手段は、前記文字校正手段により校正された後の文書から未知単語を抽出することを特徴とする実施態様１～４のいずれかに記載の情報検索装置。

【００９９】

〔実施態様７〕　前記文字校正手段は、類似の文字パターンを有する文字同士を対応付けて記憶した文字校正辞書と文字列照合のための単語辞書とを参照し、文字認識処理された文書に含まれている文字コードに前記文字校正辞書中の文字パターンを追加した文字候補列の中で、前記単語辞書に含まれる単語と一致する文字候補列を構成する文字に前記文字コードを変更することで、校正を行うことを特徴とする実施態様５または６記載の情報検索装置。

【０１００】

〔実施態様８〕　前記文書種別判別手段は、前記文書のファイル名の拡張子及び前記文書に予め付与された属性情報の少なくとも一方に基づき文書の種別を判別することを特徴とする実施態様１～７のいずれかに記載の情報検索装置。

【０１０１】

〔実施態様９〕　文書から単語を抽出する単語抽出手段と、前記単語抽出手段により抽出された単語を前記文書と対応付けて、文書検索用の索引情報として登

録する索引情報登録手段と、前記索引情報登録手段により登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索手段と、文字認識処理された文書に含まれている文字コードの誤りを校正する文字校正手段とを有し、前記単語抽出手段は、前記文字校正手段により校正された後の文書から単語を抽出することを特徴とする情報検索装置。

【０１０２】

〔実施態様１０〕　前記文字校正手段は、類似の文字パターンを有する文字同士を対応付けて記憶した文字校正辞書と文字列照合のための単語辞書とを参照し、文字認識処理された文書に含まれている文字コードに前記文字校正辞書中の文字パターンを追加した文字候補列の中で、前記単語辞書に含まれる単語と一致する文字候補列を構成する文字に前記文字コードを変更することで、校正を行うことを特徴とする実施態様９記載の情報検索装置。

【０１０３】

〔実施態様１１〕　文書から抽出された単語を前記文書と対応付けて文書検索用の索引情報として登録する索引情報登録ステップと、該索引情報登録ステップにより登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索ステップとを有する情報検索方法であって、前記文書から未知単語を抽出する未知単語抽出ステップと、前記文書の種別を判別する文書種別判別ステップと、前記文書種別判別ステップにより判別された文書の種別に応じて、前記未知単語抽出ステップにより抽出された未知単語の前記索引情報登録ステップによる索引情報としての登録の可否を決定する登録可否決定ステップとを有することを特徴とする情報検索方法。

【０１０４】

〔実施態様１２〕　文書から単語を抽出する単語抽出ステップと、前記単語抽出ステップにより抽出された単語を前記文書と対応付けて、文書検索用の索引情報として登録する索引情報登録ステップと、前記索引情報登録ステップにより登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索ステップと、文字認識処理された文書に含まれている文字コードの誤りを校正する文字校正ステップとを有し、前記単語抽出ステップは、前記文字校正

ステップにより校正された後の文書から単語を抽出することを特徴とする情報検索方法。

【０１０５】

〔実施態様１３〕　文書から抽出された単語を前記文書と対応付けて文書検索用の索引情報として登録する索引情報登録ステップと、該索引情報登録ステップにより登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索ステップとをコンピュータに実行させる情報検索プログラムであって、前記文書から未知単語を抽出する未知単語抽出ステップと、前記文書の種別を判別する文書種別判別ステップと、前記文書種別判別ステップにより判別された文書の種別に応じて、前記未知単語抽出ステップにより抽出された未知単語の前記索引情報登録ステップによる索引情報としての登録の可否を決定する登録可否決定ステップとをコンピュータに実行させることを特徴とする情報検索プログラム。

【０１０６】

〔実施態様１４〕　文書から単語を抽出する単語抽出ステップと、前記単語抽出ステップにより抽出された単語を前記文書と対応付けて、文書検索用の索引情報として登録する索引情報登録ステップと、前記索引情報登録ステップにより登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索ステップと、文字認識処理された文書に含まれている文字コードの誤りを校正する文字校正ステップとをコンピュータに実行させる情報検索プログラムであって、前記単語抽出ステップは、前記文字校正ステップにより校正された後の文書から単語を抽出することを特徴とする情報検索プログラム。

【０１０７】

〔実施態様１５〕　文書から抽出された単語を前記文書と対応付けて文書検索用の索引情報として登録する索引情報登録ステップと、該索引情報登録ステップにより登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索ステップとをコンピュータに実行させる情報検索プログラムを記憶したコンピュータ読み取り可能な記憶媒体であって、前記文書から未知単語を抽出する未知単語抽出ステップと、前記文書の種別を判別する文書種別判別ス

テップと、前記文書種別判別ステップにより判別された文書の種別に応じて、前記未知単語抽出ステップにより抽出された未知単語の前記索引情報登録ステップによる索引情報としての登録の可否を決定する登録可否決定ステップとをコンピュータに実行させる情報検索プログラムを記憶したことを特徴とする記憶媒体。

【０１０８】

〔実施態様１６〕　文書から単語を抽出する単語抽出ステップと、前記単語抽出ステップにより抽出された単語を前記文書と対応付けて、文書検索用の索引情報として登録する索引情報登録ステップと、前記索引情報登録ステップにより登録された索引情報を参照して、要求された検索用情報に対応する文書を検索する文書検索ステップと、文字認識処理された文書に含まれている文字コードの誤りを校正する文字校正ステップとをコンピュータに実行させる情報検索プログラムを記憶し、前記単語抽出ステップは、前記文字校正ステップにより校正された後の文書から単語を抽出することを特徴とする記憶媒体。

【０１０９】

【発明の効果】

以上説明したように、本発明の請求項１によれば、文書の種別に応じて未知単語の索引登録の可否を決定することで、検索漏れ及び誤検索の少ない適切な検索を行うことができる。

【０１１０】

本発明の請求項２によれば、校正後に単語抽出することで、検索精度を向上させることができる。

【図面の簡単な説明】

【図１】

本発明の一実施の形態に係る情報検索装置の全体構成を示すブロック図である。

【図２】

本発明の実施の形態の情報検索装置の機能構成を示すブロック図である。

【図３】

検索語を指示する場合において、表示器ＣＲＴに画面表示される操作パネルの

例を示す図である。

【図４】

検索要求入力処理部による検索処理実行後の検索結果の表示例を示す図である。

【図５】

文字を表す画像に対して文字認識処理を施した結果得られた抽出テキストを示す図である。

【図６】

図５に例示された文字認識誤りを含んでいる抽出テキストに対して文字校正処理を行った場合の結果を示す図である。

【図７】

文字校正用辞書の構成を示す概念図である。

【図８】

単語抽出辞書の構成を示す概念図である。

【図９】

単語インデックスの構成を示す概念図である。

【図１０】

文字校正処理で用いられる文字候補ラティスの構成を示す概念図である。

【図１１】

本実施の形態における文書検索処理のフローチャートを示す図である。

【図１２】

本実施の形態における文書登録処理のフローチャートを示す図である。

【図１３】

図１２のステップＳ３００３で実行される文字校正処理のフローチャートを示す図である。

【図１４】

図１２のステップＳ３００４で実行される単語抽出処理のフローチャートを示す図である。

【符号の説明】

１１　ＣＰＵ（索引情報登録手段、文書検索手段、未知単語抽出手段、単語抽出手段、文書種別判別手段、文字認識処理手段、登録可否決定手段、文字校正手段）

２０１　検索要求入力処理部

２０２　検索語保持部

２０３　検索処理部（文書検索手段）

２０４　検索結果保持部

２０５　単語インデックス

２０６　単語抽出処理部（未知単語抽出手段、単語抽出手段）

２０７　単語抽出辞書（単語辞書）

２０８　未知語検索指定保持部（可否指定手段）

２０９　文字校正処理部（文字校正手段）

２１０　文字校正用辞書

２１１　文字認識処理部（文字認識処理手段）

２１２　抽出テキスト

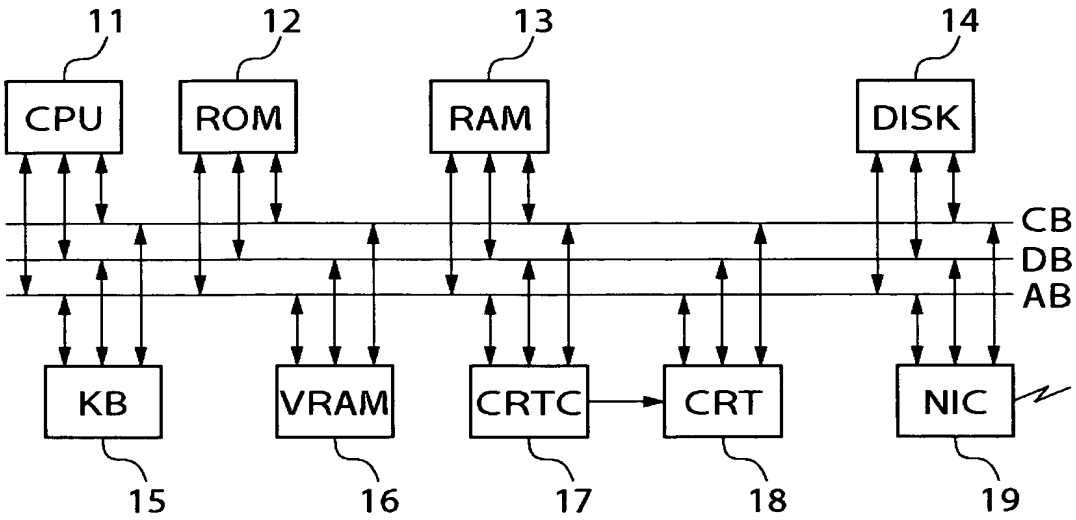
２１３　登録文書タイプ判定部（文書種別判別手段）

２１４　文書タイプ保持部

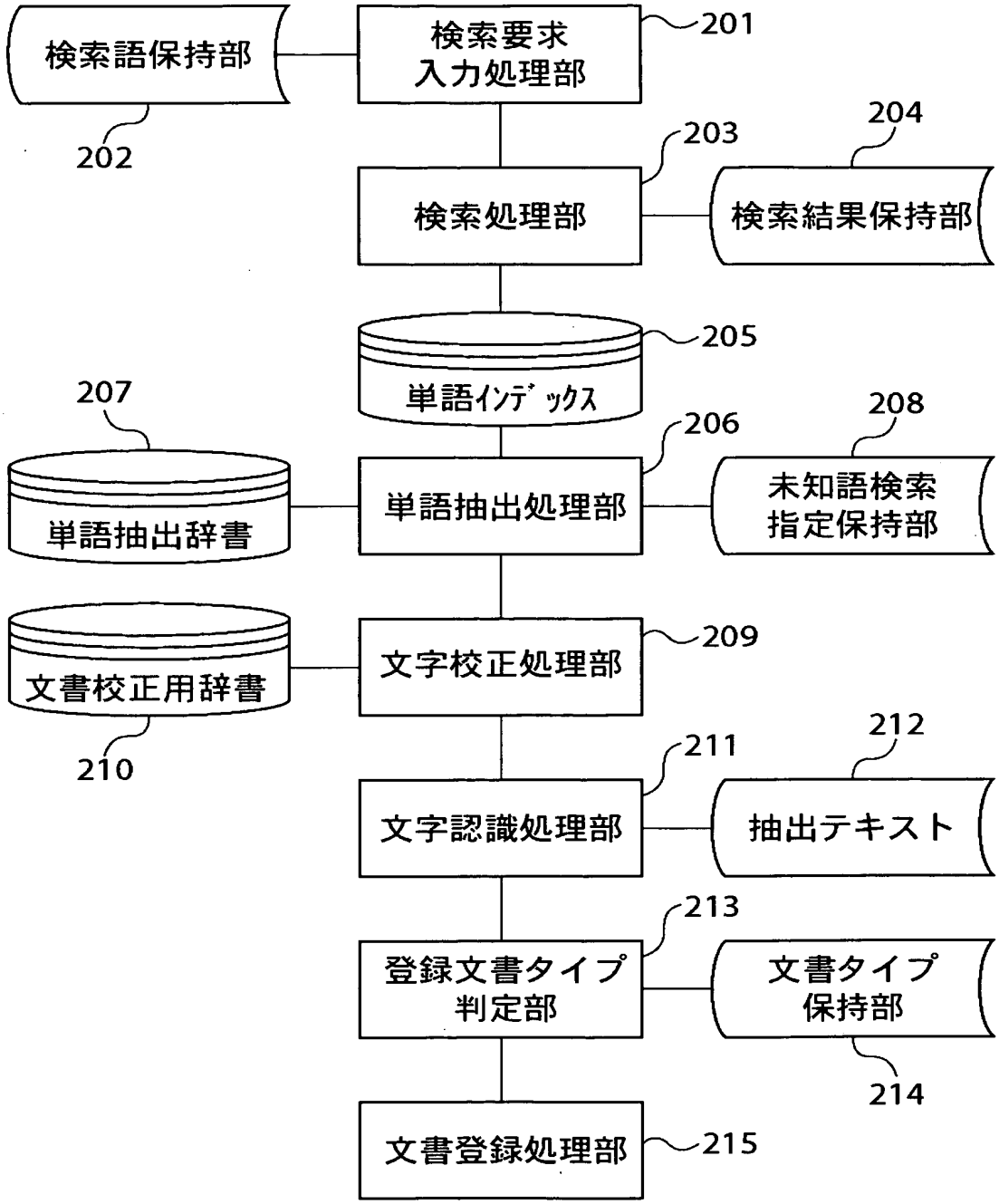
２１５　文書登録処理部

【書類名】　図面

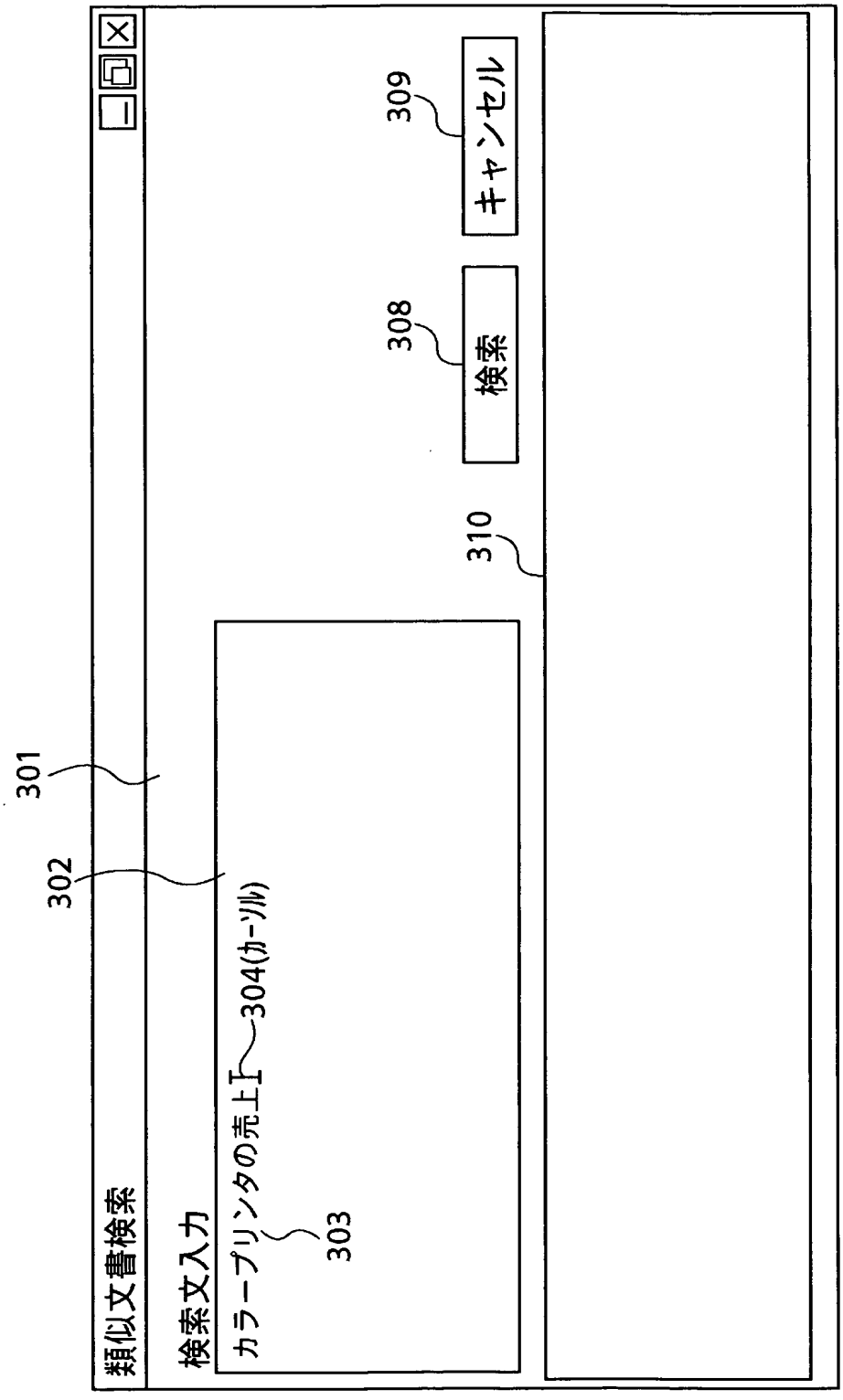
【図１】


11
12
13
14
CPU
ROM
RAM
DISK
CB
DB
AB
KB
VRAM
CRTC
CRT
NIC
15
16
17
18
19

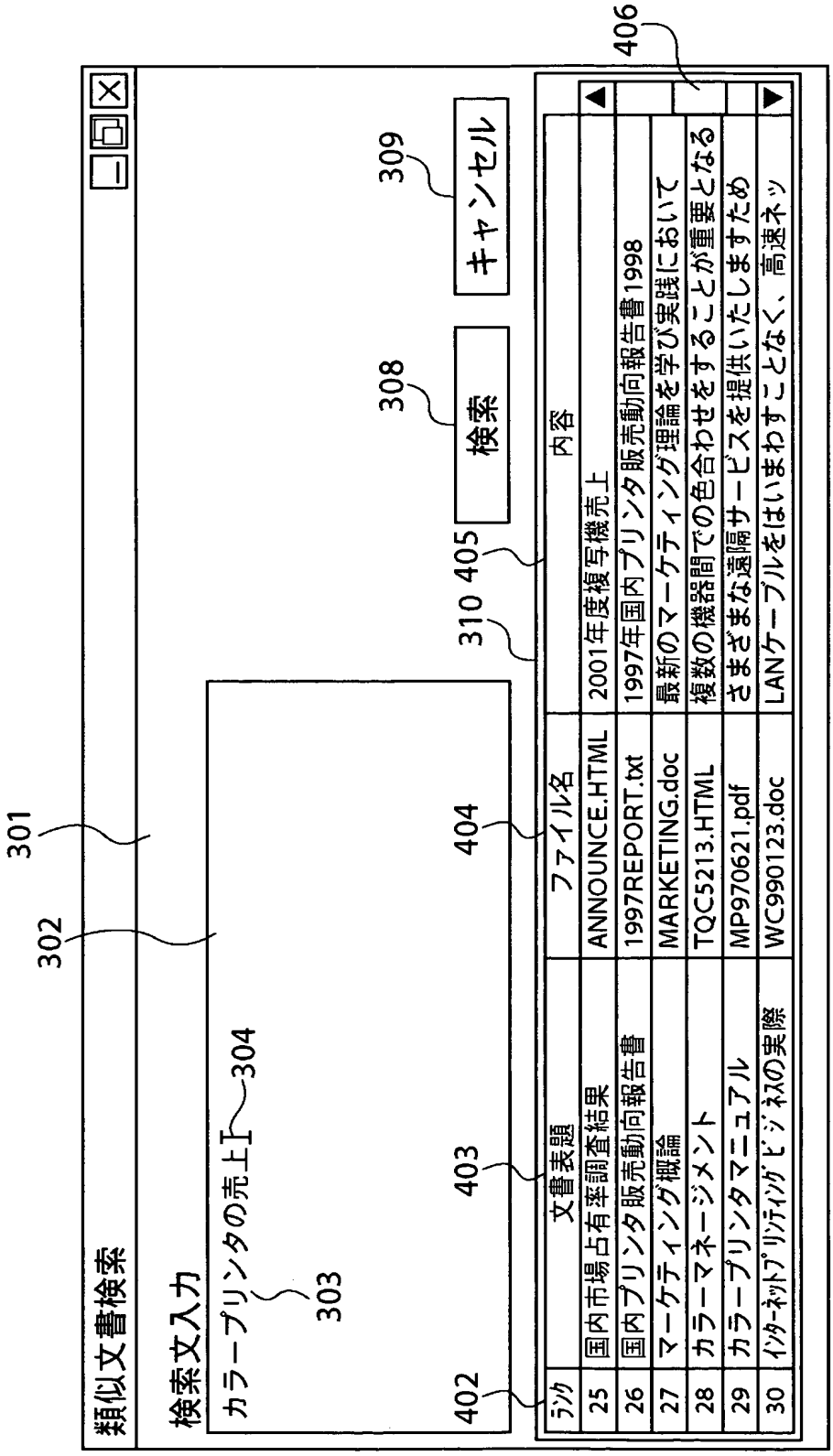
【図２】


検索語保持部
202
検索要求
入力処理部
201
検索処理部
203
検索結果保持部
204
単語ｲﾝﾃﾞｯｸｽ
205
単語抽出辞書
207
単語抽出処理部
206
未知語検索
指定保持部
208
文書校正用辞書
210
文字校正処理部
209
文字認識処理部
211
抽出テキスト
212
登録文書タイプ
判定部
213
文書タイプ
保持部
214
文書登録処理部
215

【図３】


類似文書検索
検索文入力
カラープリンタの売上
304(カーソル)
303
302
301
308
検索
309
キャンセル
310

【図４】


類似文書検索
検索文入力
カラープリンタの売上
301
302
303
304
402
403
404
405
406
308
309
310
検索
キャンセル
ランク
文書表題
ファイル名
内容
25 国内市場占有率調査結果 ANNOUNCE.HTML 2001年度複写機売上
26 国内プリンタ販売動向報告書 1997REPORT.txt 1997年国内プリンタ販売動向報告書1998
27 マーケティング概論 MARKETING.doc 最新のマーケティング理論を学び実践において
28 カラーマネージメント TQC5213.HTML 複数の機器間での色合わせをすることが重要となる
29 カラープリンタマニュアル MP970621.pdf さまざまな遠隔サービスを提供いたしますため
30 インターネットプリンティングビジネスの実際 WC990123.doc LANケーブルをはいまわすことなく、高速ネッ

【図５】

・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
このモノクロの写真に写っているモルールは当時の重要な交通手段であった・・・・・

【図６】

・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
このモノクロの写真に写っているモルールは当時の重要な交通手段であった・・・・・

【図７】

ヵカ力刀
ーーー
口口
へへ
工工
二二
……

【図８】

801　　802

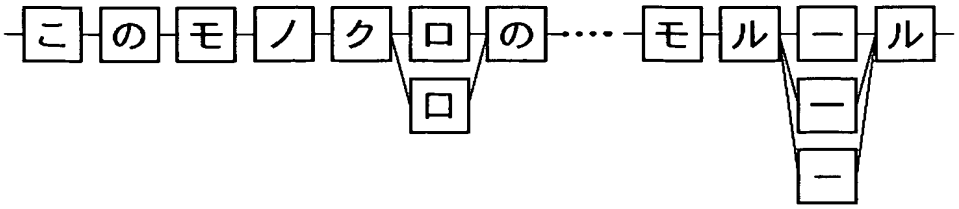
見出し語	品詞
…	…
カラー	名詞
…	…
モノクロ	名詞
…	…

【図９】

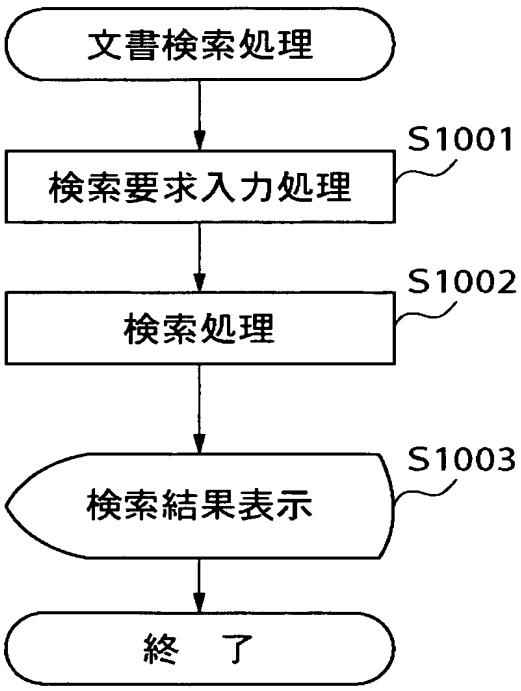
901　　902

インデックス見出し語	文書情報
…	…
カラー	(1000,15),(1200,5),(1252,8),(1800,3)
カレンダー	(328,2),(512,1),(999,2)
…	…
モノクロ	(2100,1)
…	…

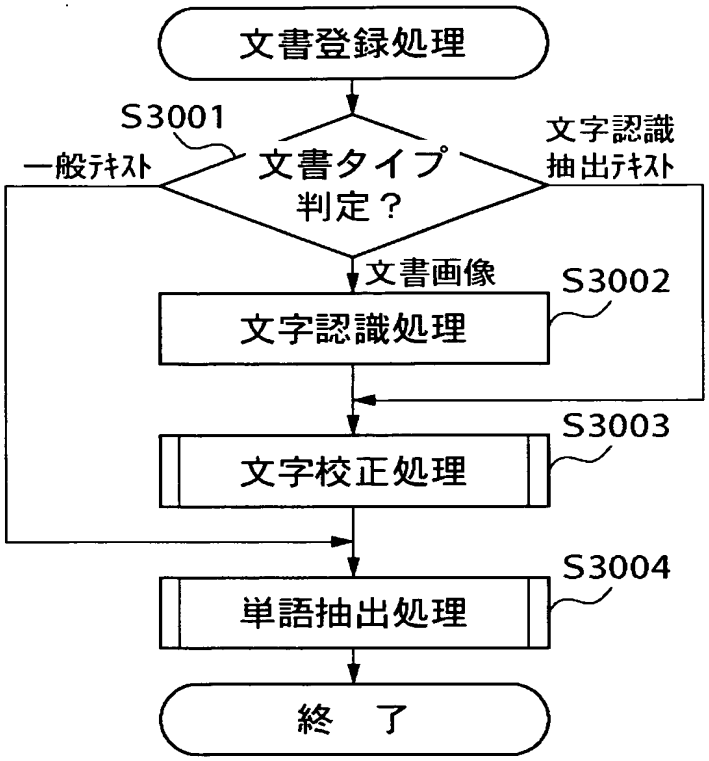
【図１０】


こ
の
モ
ノ
ク
ロ
の
ロ
…
モ
ル
ー
ル
ー
ー

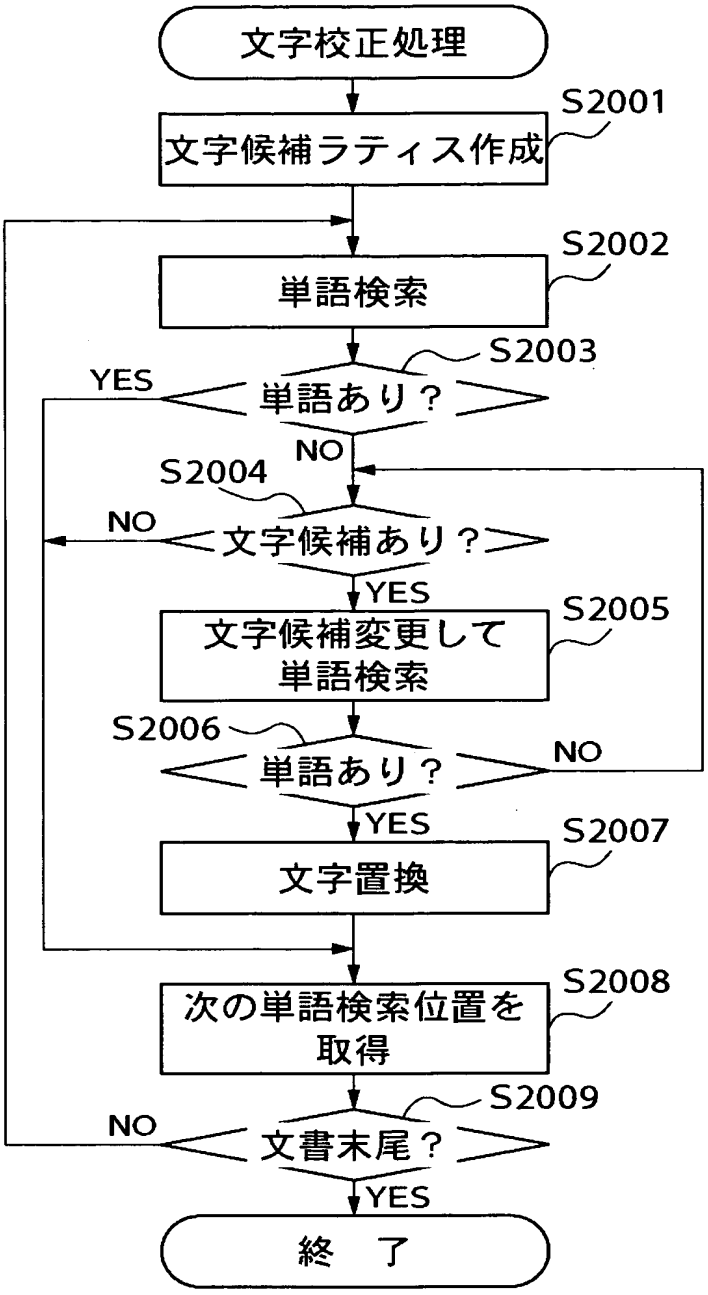
【図１１】


文書検索処理
S1001
検索要求入力処理
S1002
検索処理
S1003
検索結果表示
終　了

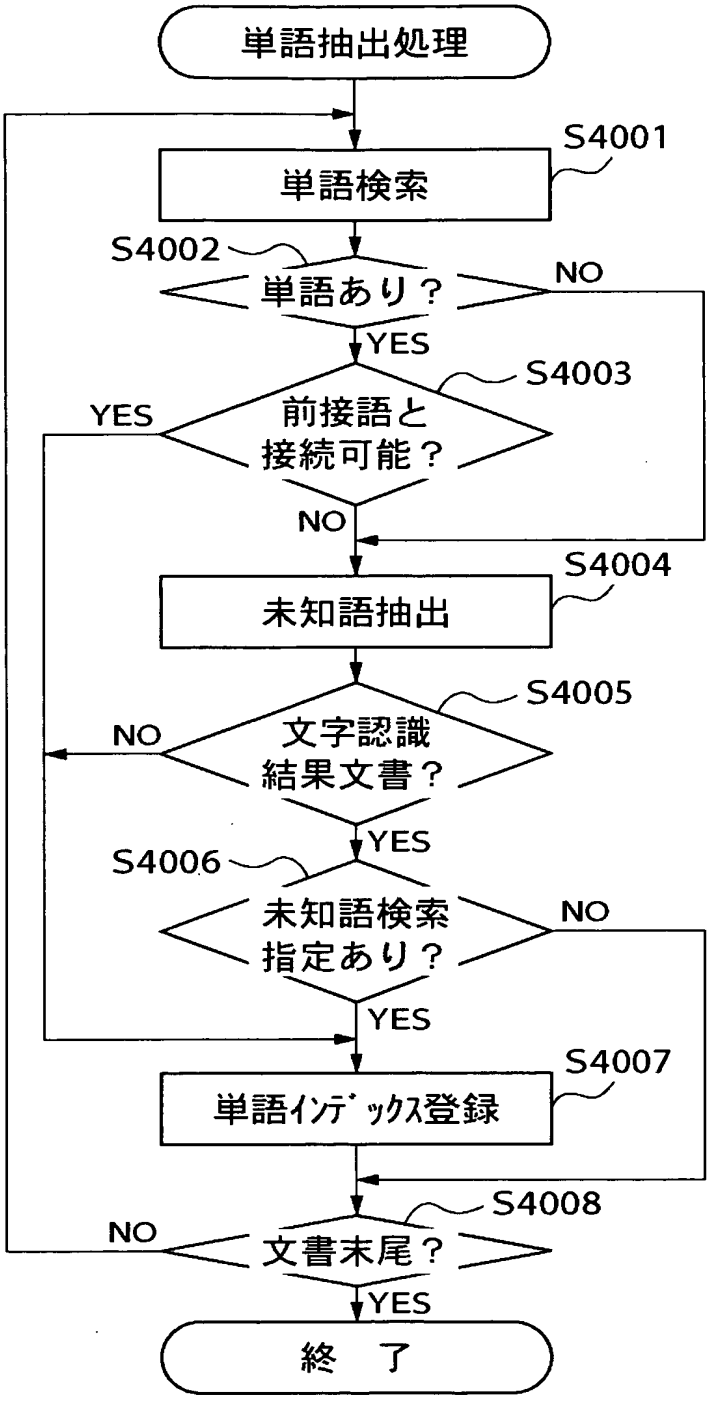
【図１２】


文書登録処理
S3001
一般ﾃｷｽﾄ
文書タイプ
判定？
文字認識
抽出ﾃｷｽﾄ
文書画像
S3002
文字認識処理
S3003
文字校正処理
S3004
単語抽出処理
終　了

【図１３】


文字校正処理
S2001
文字候補ラティス作成
S2002
単語検索
S2003
YES
単語あり？
NO
S2004
NO
文字候補あり？
YES
S2005
文字候補変更して
単語検索
S2006
単語あり？
NO
YES
S2007
文字置換
S2008
次の単語検索位置を
取得
S2009
NO
文書末尾？
YES
終　了

【図１４】


単語抽出処理
S4001
単語検索
S4002
単語あり？
NO
YES
S4003
前接語と
接続可能？
YES
NO
S4004
未知語抽出
S4005
文字認識
結果文書？
NO
YES
S4006
未知語検索
指定あり？
NO
YES
S4007
単語ｲﾝﾃﾞｯｸｽ登録
S4008
文書末尾？
NO
YES
終　了

【書類名】　要約書

【要約】

【課題】　検索漏れ及び誤検索の少ない適切な検索を行えるようにする。

【解決手段】　単語抽出処理対象の文書から単語検索、または辞書登録されていない未知語の抽出を行い、単語抽出処理対象文書が文字認識結果文書、すなわち、文字認識処理済み文書でない場合は、単語インデックス２０５に索引情報として登録する。一方、文字認識処理済み文書である場合は、未知語検索指定保持部２０８を参照することで、未知語検索指定の有無を判定する。そして、「未知語検索指定有り」である場合は、抽出された未知語を、単語インデックス２０５に索引情報として登録する一方、「未知語検索指定無し」である場合は、索引情報として登録しない。

【選択図】　図１４

特願２００３－０１３４２８

出　願　人　履　歴　情　報

識別番号　［０００００１００７］

１．変更年月日　１９９０年　８月３０日
［変更理由］　新規登録
住　所　東京都大田区下丸子３丁目３０番２号
氏　名　キヤノン株式会社